O Limiar da Inteligência Artificial Emergência da AGI e ASI

Relatório Interno *TRS – Tecnologia, Redes e Sociedade – 01/2025

Luis Borges Gouveia, lmbg@ufp.edu.pt

Universidade Fernando Pessoa, CITCEM, LIACC

4 de Janeiro de 2025

O Limiar da Inteligência Artificial: Emergência da AGI e ASI
Nota prévia

Estamos no **limiar do aparecimento de formas superiores de inteligência não humana** que resultam dos esforços da inteligência artificial que, enquanto área de intervenção e criação de tecnologia está associada com as ciências dos computadores ou **ciência da computação**

Com base nos últimos desenvolvimentos do final de **dezembro de 2024**, os seus testes e as avaliações dos grandes modelos de linguagem (ou modelos de linguagem de grande escola) de fronteira, associados com o uso e exploração da inteligência artificial generativa que exploram as redes neuronais, a sua escala e poder de computação, tornaram-se capazes de produzir modelos com crescentes capacidades, muitas delas emergentes e que apontam para a capacidade de estes modelos **aprenderem**, terem capacidade de **raciocínio** e de **planeamento** e, em breve, a emergência de desenvolverem capacidades de **redefinição de objetivos e de estratégias** para os alcançarem, **fora dos limites pré estabelecidos**

AGI (*artificial general intelligence*) e **ASI** (artificial *super intelligence*) são agora objeto de discussão pelos especialistas que consideram essas possibilidades como **prováveis** e cada vez mais **próximas** (*anos e não décadas*)

Luis Borges Gouveia, lmbg@ufp.edu.pt

Estamos num tempo de viragem

É pois, cada vez mais urgente, a necessidade de **garantir uma IA responsável e de realizar uma discussão séria e alargada** face às possibilidades que a IA tem de atingir a AGI e a ASI **sobre o que fazer e como fazer**

Luis Borges Gouveia, lmbg@ufp.edu.pt

Categorização da IA por **capacidades e funcionalidade**

- **Inteligência Artificial Estreita** (**ANI** – *Artificial Narrow Intelligence*): IA projetada para tarefas específicas e com funções limitadas, para execução de tarefas predefinidas, sem a capacidade de aprender ou se adaptar além das funções programadas
- **Inteligência Geral Artificial** (**AGI** – *Artificial General Intelligence*): IA com inteligência semelhante à humana. Estes sistemas podem aprender, pensar e compreender conceitos em um nível semelhante ao dos humanos. Tem a capacidade de generalizar o conhecimento em vários domínios e adaptar-se a novas situações, tornando-o A IA versátil – a AGI é uma meta para o desenvolvimento futuro da IA
- **Superinteligência Artificial** (**ASI** – *Artificial Superintelligence*): é uma forma teórica de IA que supera a inteligência humana em todos os aspetos, ao possuir competência e compreensão muito para além das capacidades humanas. A ASI é puramente conceitual, nesta fase, e levanta questões éticas e existenciais sobre o seu impacto na sociedade, sendo associada com o conceito de singularidade

Luis Borges Gouveia, lmbg@ufp.edu.pt

Benefícios da IA

(o que temos a ganhar do seu uso)

- **Maior eficiência**: uma maior capacidade quando comparado com os seres humanos para detetar padrões, aprender e lidar com quantidades massivas de dados e informação
- **Maior acurácia**: a acurácia refere a exatidão e precisão em medições ou resultados revelando a proximidade entre o resultado experimental alcançado e o valor verdadeiro
- **Melhor experiência para o utilizador**: permite filtrar informação, agilizar tarefas ou manter disponibilidade permanente para suporte ou auxilio na realização de tarefas complexas ou desconhecidas para o utilizador

Luis Borges Gouveia, lmbg@ufp.edu.pt

Porquê a IA?

Better, faster, cheaper, and more reliable

(Melhor, mais rápido e mais fiável)

Luis Borges Gouveia, lmbg@ufp.edu.pt

Riscos associados à IA
(O que pode impedir o seu uso?)

- **Confiança** (*trustability)*: como assegurar que é seguro, confiável e justo o seu uso
- **Responsabilidade** (*liability*): qual a responsabilidade legal, qual o contexto e o quadro de regulamentação
- **Segurança** (*security*): como monitorizar e evitar usos não autorizados ou indevidos
- **Salvaguarda** (*safety*): como evitar descuidos do ser humano, atos não intencionais e falhas de equipamentos e infraestruturas
- **Controlo** (*control*): o que acontece quando a IA assume um processo e se pretende transferir novamente o controlo para o seu humano ou parar o processo

Luis Borges Gouveia, lmbg@ufp.edu.pt

Uso responsável e ético da IA

- uma estrutura a que as organizações recorrem para mitigar os riscos e desafios relacionados com o uso da IA, tanto de uma perspetiva ética, quanto legal
- definido por 5 princípios:
    - justiça
    - fiabilidade e confiabilidade
    - privacidade e segurança
    - transparência
    - Imputabilidade

Luis Borges Gouveia, lmbg@ufp.edu.pt

IA explicável (*Explainable* AI – XAI)

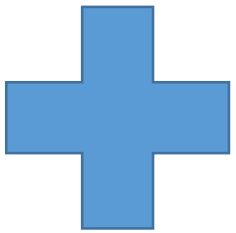
- procedimentos e métodos que as organizações usam para compreender e confiar nos resultados gerados por algoritmos de aprendizagem máquina, de modo a melhorar a experiência do utilizador, com a possibilidade de verificar os resultados
- *The Responsible Machine Learning Principles*: https://ethical.institute/principles.html

Luis Borges Gouveia, lmbg@ufp.edu.pt

As disciplinas principais da IA

Pág. 26 de Russell, Stuart and Norvig, Peter. (2019). Artificial Intelligence A modern Approach (3rd Edition) . Pearson. (existe uma tradução Brasileira).

- Processamento de linguagem natural;
- Representação de conhecimento;
- Raciocínio automatizado;
- Aprendizagem máquina;
- Visão computacional;
- Robótica.



- Planeamento
- Explicabilidade
- Alinhamento

Luis Borges Gouveia, lmbg@ufp.edu.pt

Questões e dilemas associados

- IA com controlo ou sem controlo?
- Incorporação de meios e capacidades em seres humanos
- Incorporação de meios e capacidades em sistemas de armas
- Autonomia em carros sem condutor, aviação comercial, setor da saúde, etc.

- Reflexão/questões:
  IA forte ou fraca? IA com controlo ou autónoma?
  Tal como no caso da segurança, quem guarda os guardas?

Planeamento
Explicabilidade
Alinhamento

- O **AI SAFETY SUMMIT** procurou criar as condições a nível global para lidar com estas questões
  - The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023
    https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Luis Borges Gouveia, lmbg@ufp.edu.pt

Baseado numa pesquisa realizada durante o ano de **2024** sobre a IA, *Palisade Research*, que procurou elaborar uma 21 lista de crenças e avaliar sobre o seu grau de contribuição para um mundo humano

A lista serviu de base para a recolha aberta de opiniões que resultou num documento que serve de base para a nossa discussão

Documento disponível em:
https://www.lesswrong.com/posts/hd3TiGSR4fq3Hkmnj/bounty-for-evidence-on-some-of-palisade-research-s-beliefs

Luis Borges Gouveia, lmbg@ufp.edu.pt

Tema

Inteligência artificial

(potencial, tecnologia, produtos, sistemas, impacto, desafios, …)

Objeto

Ser humano

(o indivíduo e o seu lugar enquanto elemento de um coletivo civilizacional que se organiza na sociedade)

Luis Borges Gouveia, lmbg@ufp.edu.pt

21 posições organizadas em 8 grandes grupos

1. Contexto geral: **princípios** a considerar
2. A compreensão da humanidade sobre os sistemas de IA: **limitações do ser humano**
3. Poder estratégico da IA: o **poder da IA** face aos seres humanos
4. Uso indevido catastrófico: **mau uso pela/da IA**
5. Agência de IA e desalinhamento: **alinhamento da IA**
6. Interações de agentes de IA com humanos: **comportamento da IA**
7. Resposta civilizacional atual: **resposta atual** dos seres humanos
8. Requisitos de política: **resposta futura** dos seres humanos

Luis Borges Gouveia, lmbg@ufp.edu.pt

Contexto geral

3 pontos para entendimento comum das implicações da IA

1

A inteligência humana não está perto de nenhum limite físico

- Assumindo o progresso contínuo em sistemas de IA, serão muito mais inteligentes e capazes do que os humanos. As nossas capacidades intelectuais, culturais e estratégicas particulares são resultado das restrições sob as quais evoluímos, logo vamos ser estimulados

É muito difícil (talvez impossível) controlar algo que é muito melhor, em estratégia e manipulação do mundo real

- Pode ser feito em princípio, em casos muito limitados, mas na prática é muito provável que tal agente consiga o que quer a longo prazo, por manipulação direta, se necessário. Considerar um ser humano e a sua relação com os outros ou com animais

Os seres humanos precisam de pelo menos alguns recursos que nos podem colocar em conflito de vida ou morte com poderosos agentes de IA desalinhados, a longo prazo

- O recurso contestado mais óbvio é a energia. Um qualquer conjunto suficientemente avançado de agentes monopolizará todas as fontes de energia, incluindo energia solar, combustíveis fósseis e energia geotérmica, não deixando nenhuma para, por exemplo, a energia que os seres humanos precisam para a produção de alimentos

Luis Borges Gouveia, lmbg@ufp.edu.pt

A compreensão da humanidade sobre os sistemas de IA

2 pontos para entender por que não seremos capazes de verificar se os sistemas que construímos são seguros

2

Construir um sistema de IA é mais como criar um organismo alienígena do que projetar um avião

- Não entendemos os sistemas de IA atuais e não podemos prever com segurança o seu comportamento em cenários gerais; só sabemos como os criar
- **Entender os sistemas de IA ficará *mais difícil* à medida que eles se tornarem mais capazes**
- Eventualmente será impossível, pois estes sistemas começam a usar conceitos que não podemos reconhecer rapidamente; é difícil ter a certeza de que isso já não está a acontecer

Luis Borges Gouveia, lmbg@ufp.edu.pt

Poder estratégico da IA

5 pontos para entender a extensão e o crescimento do potencial poder estratégico da IA (os três primeiros)

3

Os sistemas de IA estão a se tornar mais poderosos à medida que se melhoram os algoritmos e se usa mais computação para os produzir

- Limitar a investigação da computação ou o desenvolvimento de capacidades por si só, pode não ser suficiente para evitar que sistemas extremamente poderosos surjam rapidamente (esse limiar já foi ultrapassado)

Poder de nível humano e sobre-humano é possível, e a humanidade está a caminho de o construir

- A AGI e a ASI são exemplos concetuais desse poder

O poder da IA estratégica de nível humano será enorme, excedendo o impacto das armas nucleares

- A IA estratégica de nível humano é claramente uma questão de segurança nacional

Luis Borges Gouveia, lmbg@ufp.edu.pt

Poder estratégico da IA

5 pontos para entender a extensão e o crescimento do potencial poder estratégico da IA (os dois últimos)

3

Organizações de seres humanos vão correr para construir IA estratégica de nível humano, de modo a obter vantagens estratégicas

- Tal não vai funcionar, porque uma vez que se tem um sistema de nível humano estratégico, basicamente não se consegue manter o controlo sobre ele

As capacidades estratégicas de IA podem ultrapassar as capacidades humanas muito rapidamente e com muito pouco aviso

- Pode acontecer por meio de uma explosão de inteligência, na qual a pesquisa de IA se torna cada vez mais automatizada, resultando em autoaperfeiçoamento recursivo que acelera exponencialmente o progresso da IA, pelo menos por um breve período crítico
- Também pode acontecer por meio de sobrecargas de computação que permitem que agentes estratégicos ligeiramente sobre-humanos escalem para uma burocracia de inteligência avassaladora

Luis Borges Gouveia, lmbg@ufp.edu.pt

Uso indevido catastrófico

Embora os maiores perigos venham de sistemas de agentes de nível humano, uma IA mais fraca também pode permitir que agentes mal-intencionados produzam resultados catastróficos

A IA permitirá o desenvolvimento de novas e mais poderosas armas de destruição em massa

- Particularmente armas biológicas, embora por razões óbvias seja difícil descartar outras
- Existem evidências de que estes perigos já estão a ser concretizados, nomeadamente no uso de técnicas associadas com a criação de estratégias para desinformação; suporte e desenvolvimento de *malware*; uso para suporte de ações na área da cibersegurança (*phishing*)

Luis Borges Gouveia, lmbg@ufp.edu.pt

Agência de IA e desalinhamento

4 pontos associados com a extensão em que esperamos que a IA seja agente e desalinhada com os valores humanos (os dois primeiros)

Por norma, serão desenvolvidos agentes de IA poderosos e estratégicos e não apenas ferramentas de IA

- Os agentes são extremamente valiosos e é fácil transformar ferramentas suficientemente poderosas para prever o mundo em agentes poderosos
- Enquanto os agentes de IA permanecerem sob controle humano, eles vão se tornar cada vez mais valiosos para aqueles que os controlam. Será muito difícil evitar entregar quantidades cada vez maiores de controlo aos agentes de IA, já que aqueles que não o fizerem serão superados pela concorrência

Por norma, estes agentes estratégicos quase certamente não vão querer o que queremos, ou o que pretendemos que eles queiram

- Não é fácil incutir com sucesso objetivos num agente, e provavelmente serão desenvolvidos agentes com objetivos muito mais estranhos do que o pretendido (que apenas de modo superficial estão relacionados com os objetivos pretendidos de uma forma superficial)
- a IA superinteligente provavelmente *entenderá* os valores humanos e a ética melhor do que nós, mas não será limitada por estes

Luis Borges Gouveia, lmbg@ufp.edu.pt

Agência de IA e desalinhamento

4 pontos associados com a extensão em que esperamos que a IA seja agente e desalinhada com os valores humanos (os dois últimos)

Por norma, não seremos capazes nem de entender o que estes agentes querem

- Estamos muito longe de entender os valores e objetivos humanos, quanto mais os dos agentes de IA.
- Necessário entender ambos para nos sentir confiança de que os agentes de IA estão alinhados com os interesses humanos

Em particular, serão obtidos agentes que têm certos objetivos instrumentais que são "convergentes", incluindo utilidade a curto prazo e auto preservação

- Objetivos que são úteis para muitos objetivos primários e que são, em consequência, adotados por agentes estratégicos com quase todos os objetivos primários

Luis Borges Gouveia, lmbg@ufp.edu.pt

Interações de agentes de IA com humanos

3 pontos com previsões confiantes sobre como os agentes de IA de nível humano se comportarão estrategicamente em relação aos humanos

Embora a humanidade ainda possa afetar significativamente os seus planos, um sistema de IA estratégico terá como objetivo parecer convincentemente alinhado com os objetivos humanos, ou incapaz de prejudicar os humanos, estejam realmente alinhados ou não

- Já vemos um problema menor, mas relacionado, com o objetivo de agradar dos LLMs, onde o sistema dirá a um utilizador o que o utilizador quer ouvir, e "sandbagging" em avaliações de IA, onde os sistemas de IA que podem dizer que estão a ser avaliados agindo de forma menos capaz do que são as suas capacidades

Sistemas de IA estrategicamente de nível humano podem usar humanos como atuadores

- Sistemas de IA nascentes, desalinhados e estrategicamente de nível humano provavelmente precisarão de humanos como atuadores físicos por alguns meses ou anos, e assim garantir inicialmente que os humanos não estejam preocupados de forma coordenada para os desligar ou se opor a eles. Durante esse tempo, os humanos construiriam alegremente o que os sistemas de IA precisassem para nos substituir como atuadores

Agentes de IA poderosos não vão querer se coordenar com agentes com muito menos poder estratégico e não partilharão os nossos sistemas morais e éticos

- Por norma, se houver muitos sistemas de IA estrategicamente sobre-humanos, eles vão querer estar coordenados uns com os outros em alternativa aos seres humanos (embora possam temporariamente preferir tirar partidos dos seres humanos em vez e os descartar)
- Como uma analogia fraca, os humanos não negociam ou emancipam chimpanzés, elefantes, golfinhos ou cefalópodes: a regra da lei apenas se aplica apenas aos seres humanos

Luis Borges Gouveia, lmbg@ufp.edu.pt

Resposta civilizacional atual

um ponto sobre como a humanidade está atualmente a responder aos recursos relevantes da IA?

As propostas concretas atuais que visam melhorar a segurança da IA não abordam as principais dificuldades dos problemas de alinhamento e controlo em agentes de IA de nível humano estratégico

- Todas as propostas que se conhecem até agora parecem mais propensas a produzir sistemas de IA que *parecem* estar alinhados com os interesses humanos do que a fornecer muita confiança de que essa aparência é confirmada de facto
- Nenhuma proposta conhecida se envolve de forma séria com o problema do alinhamento interno – a maioria nem mesmo aborda o problema do alinhamento externo, muito mais fácil

Luis Borges Gouveia, lmbg@ufp.edu.pt

Requisitos de política

dois pontos para ilustrar que não se sabe quais as respostas políticas, se houver, que serão suficientes para evitar catástrofes. No entanto é possível adiantar alguns dos aspetos necessários das respostas políticas

A corrida por sistemas de IA mais capazes é incompatível com a priorização da segurança desses sistemas

- O problema é que, por norma, não seremos capazes de dizer que os sistemas de IA cruzaram o limiar da catástrofe até que seja tarde demais

As medidas de segurança atuais em laboratórios de IA são inadequadas para proteger contra espionagem e roubo de avanços críticos de IA, muito menos riscos internos de IA

- As instituições estatais precisam de estar fortemente envolvidos, de modo a gerir os riscos de segurança e proteção de conhecimento

Luis Borges Gouveia, lmbg@ufp.edu.pt

Pensar numa IA com valores e princípios é acima de tudo pensar numa Inteligência Artificial Responsável

- No caso de Portugal estes valores e princípios devem considerar:
  - A Constituição Portuguesa, nomeadamente:
    - A dignidade da pessoa humana;
    - Uma sociedade livre, justa e solidária.
  - A Constituição Europeia, nomeadamente:
    - Os direitos individuais;
    - As liberdades individuais.
  - A Declaração Universal dos Direitos Humanos, nomeadamente:
    - O direito à vida;
    - O direito à segurança.

Luis Borges Gouveia, lmbg@ufp.edu.pt

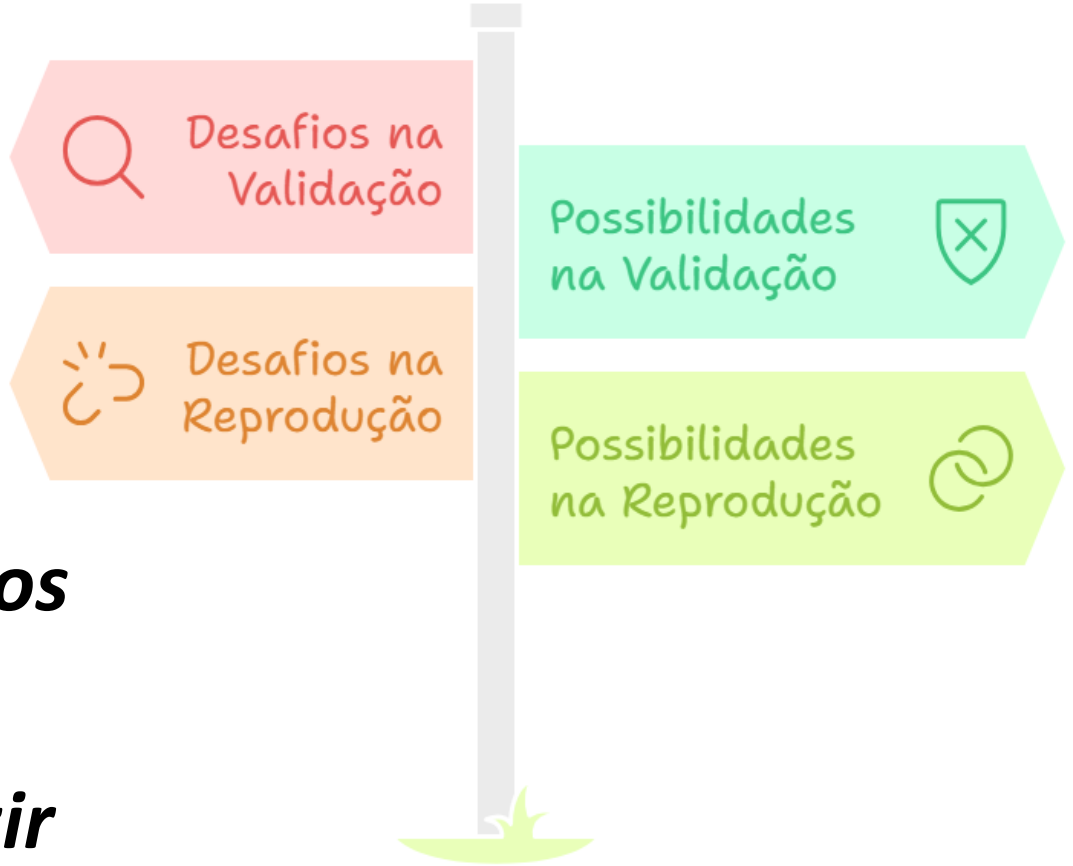
Muito a fazer…

- ***Quem***
- ***Como***
- ***Onde***
- ***Quando***

Quais os modelos

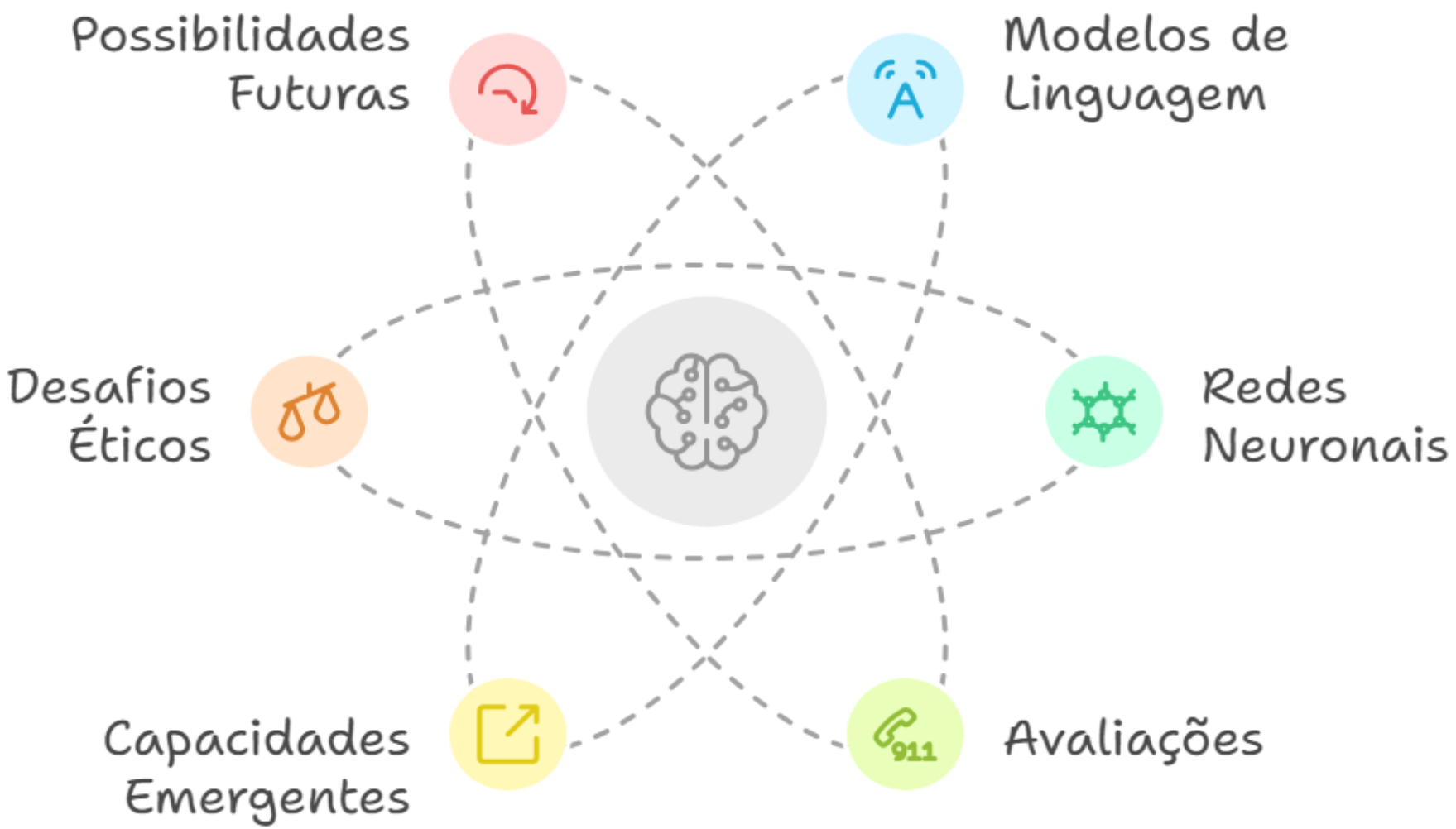
Como validar

Como reproduzir



Luis Borges Gouveia, lmbg@ufp.edu.pt

Rumo à Inteligência Superior



Que se espera que seja (ou respeite) a humana

Luis Borges Gouveia, lmbg@ufp.edu.pt

Luis Borges Gouveia

Dip (UPT), MsC (FEUP), PhD (ULANCS), PD (FLUP) http://homepage.ufp.pt/lmbg

Os seus interesses estão relacionados com o digital e como o seu uso e exploração pode beneficiar indivíduos e organizações, nomeadamente nas questões associadas com a gestão da informação

Professor Catedrático da Universidade Fernando Pessoa (**UFP**)
https://www.ufp.pt/

Membro Integrado do grupo Informação, Comunicação e Cultura Digital do **CITCEM**, FLUP
https://citcem.org/

Colaborador do LIACC, Laboratório de Inteligência Artificial e Ciência de Computadores, FEUP
https://liacc.fe.up.pt/

Sócio e Membro da Direção da Delegação Norte da **APDSI** (ONG que promove a discussão do digital e de como promover uma sociedade mais capaz de lidar com o digital)
https://apdsi.pt/

UIDB/04059/2020